研究与开发

基于视频时域感知特性的恰可察觉失真模型

邢亚芬[1]，殷海兵[1]，王鸿奎[1,2]，骆琼华[1]
（1. 杭州电子科技大学通信工程学院，浙江 杭州 310018;
2. 华中科技大学电子信息与通信学院，湖北 武汉 430074）

摘　要：现有的时域恰可察觉失真（just noticeable distortion，JND）模型对时域特征参量的作用刻画尚不够充分，导致空时域 JND 模型精度不够理想。针对此问题，提出能准确刻画视频时域特性的特征参量以及异质特征参量同质化融合方法，并基于此改进时域 JND 模型。关注前景/背景运动、时域持续时间、时域预测残差波动强度、帧间预测残差等特征参量，用来刻画视频内容的时域特征；基于人眼视觉系统（human visual system，HVS）特性探索感知概率密度函数，将异质特征参量统一映射到自信息和信息熵尺度上，实现同质化融合度量；从能量分配的角度探究视觉注意与掩蔽的耦合方法，并据此构建时域 JND 权重模型。在空域 JND 阈值的基础上，融合时域权重以得到更加准确的空时域 JND 模型。为了评估空时域 JND 模型的性能，进行了主观质量评估实验，与现有的 JND 模型相比，在感知质量接近的情况下，提出的空时域 JND 模型能够容忍更多失真，具有更强的掩藏噪声的能力。

关键词：恰可察觉失真；人眼视觉特性；视觉掩蔽；视觉注意；自信息；信息熵
中图分类号：TN919
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2022030

Video temporal perception characteristics based just noticeable difference model

XING Yafen[1], YIN Haibing[1], WANG Hongkui[2], LUO Qionghua[1]
1.College of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China
2.College of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China

Abstract: The existing temporal domain JND(just noticeable distortion) models are not sufficient to depict the interaction between temporal parameters and HVS characteristics, leading to insufficient accuracy of the spatial-temporal JND model. To solve this problem, feature parameters that can accurately describe the temporal characteristics of the video were explored and extracted, as well as a homogenization method for fusing heterogeneous feature parameters, and the temporal domain JND model based on this was improved. The feature parameters were investigated including foreground and background motion, temporal duration along the motion trajectory, residual fluctuation intensity along

收稿日期：2021−07−16；修回日期：2021−12−22
基金项目：国家自然科学基金资助项目（No.61972123, No.61931008, No.62031009）；浙江省“尖兵”研发攻关计划项目（No.2022C01068）；
Foundation Items: The National Natural Science Foundation of China (No.61972123, No.61931008, No.62031009), Zhejiang Provincial Vanguard Research and Development Project (No.2022C01068)

motion trajectory and adjacent inter-frame prediction residual, etc., which were used to characterize the temporal characteristics. Probability density functions for these feature parameters in the perception sense according to the HVS(human visual system) characteristics were proposed, and uniformly mapping the heterogeneous feature parameters to the scales of self-information and information entropy to achieve a homogeneous fusion measurement. The coupling method of visual attention and masking was explored from the perspective of energy distribution, and the temporal-domain JND weight model was constructed accordingly. On the basis of the spatial JND threshold, the temporal domain weights was integrated to develop a more accurate spatial-temporal JND model. In order to evaluate the performance of the spatiotemporal JND model, a subjective quality evaluation experiment was conducted. Experimental results justify the effectiveness of the proposed model.

Key words: JND, HVS characteristics, visual masking, visual attention, self-information, information entropy

0　引言

人眼视觉系统（human visual system，HVS）的敏感度有限，只能察觉图像大于某一阈值的失真，这个阈值被称为恰可察觉失真（just noticeable distortion，JND）阈值。由于 JND 描述了 HVS 在视觉内容上的可见性，因此在图像和视频处理领域，如果可以充分利用 HVS 特性进行压缩，能够在保证视频主观质量的情况下显著减少视频传输的码率。因此，基于 HVS 特性建模的 JND 阈值被广泛用于优化量化、运动估计以及视频目标质量评价等[1]。

视频中的特征参量对人眼视觉感知有重要影响。Korhonen[2]提取视频中 74 个特征参量用于视频质量评价。传统 JND 建模实验根据不同视频内容，如空域的亮度、边缘、纹理、颜色等，时域视频特征如目标运动、时域频率等，考虑了一些主要的 HVS 特性，如亮度自适应（luminance adaption，LA）、对比掩蔽（contrast masking，CM）效应、中心凹掩蔽（foveated masking，FM）效应、空时域对比敏感函数（contrast sensitivity function，CSF）、时域掩蔽（temporal masking，TM）效应、视觉注意特性等。现有的 JND 模型根据计算域不同一般可以分为两类：像素域 JND 模型和基于子带（DCT 或小波变换）的 JND 模型。像素域 JND 模型可以直接计算，因此时域 JND 建模在像素域主要利用相邻帧间亮度差，如 Chou 等人[3]和 Yang 等人[4]依据相邻帧的平均亮度差作为时域的不连续性构建时域 JND 模型，Chin 等人[5]利用连续帧的运动变化估计 JND 阈值。时域基于子带的 JND 建模主要考虑 CSF 和人眼运动特性：Kelly 等人[6]依据收集的视网膜稳定行波刺激实验的数据，首次提出了空时域 CSF 模型。之后，Daly 等人[7]加入人眼运动特性对 Kelly 的模型做了补充。Jia 等人[8]在他们的基础上将空时域 CSF 和人眼的运动特性相结合。此外，Wei 等人[9]考虑人眼对不同运动方向的敏感程度，构建了适合于单通道灰度视频的变换域 JND 模型。Bae 等人[10]将时域掩蔽效应和中心凹掩蔽效应结合起来，提出一种时域中心凹掩蔽（temporal-foveated masking，TFM）模型，估计运动目标的 JND 阈值。此外，由于机器学习在感知领域的广泛应用，Wang 等人[11-12]构建了两个基于 JND 压缩视频质量的大规模数据集：MCL-JCV 和 VideoSet。Ki 等人[13]提出基于线性回归的 LR-JNQD 模型和基于卷积神经网络的 CNN-JNQD 模型。Liu 等人[14]提出图像的 PW-JND 预测模型，基于深度学习的二元分类器能够根据图像感知质量是否有损来决定图像的 JND 值。Zhang 等人[15]提出基于深度学习的视频感知质量预测模型，它能够预测不同分辨率和编码参数的压缩视频的感知质量，从而计算 JND 的阈值。

然而，由于视频时域场景的复杂和多样性，传统方法对时域视觉特征参量的提取并不充分，

导致时域仍有大量冗余信息。因此，若需要准确估计视频时域 JND 阈值，就需要分析时域不同特征参量对人眼视觉感知的影响。Wang 等人[16]研究了视频中普遍存在的 3 种运动——绝对运动、相对运动和背景运动，并分析由这 3 种运动引起的视觉注意和掩蔽效应，提出了基于运动感知的视频质量评价方法。基于这个启发，分析视频中影响人眼视觉感知特性的 5 个特征参量，考虑利用相对运动和运动轨迹上的持续时间度量视觉显著度，由背景运动、帧间残差波动强度和相邻帧间预测残差度量不确定度，时域感知特征参量提取方法如图 1 所示。这些特征参量在一定程度上会对人眼感知失真造成影响。一方面，部分激励源会影响人眼关注的定睛点，另一方面，有些特征参量会降低视觉感知灵敏度，消耗人眼感知能量。这些不同激励之间的相互作用或相互干扰就造成了视觉掩蔽效应[17]。探究这些激励的耦合作用一直以来是 JND 建模重要的研究内容之一[16]。例如像素域建模时普遍采用非线性相加模型（nonlinear additivity model for masking，NAMM），能够消除掩蔽效应的重叠部分，而 DCT 域 JND 一般建模为多个掩蔽因子的乘积。但这些模型或 HVS 特性考虑不全面导致计算精度不足，或者计算方式复杂导致难以融合，以致 JND 阈值估计不准确。因此，提出一种统一有效的感知特征参量度量方法以对异质感知特征参量进行同化是 JND 建模过程中需要解决的重要问题，若能以相同的标准去度量这些特征参量，有利于更准确地分析不同视觉特征之间的相互耦合作用。

一般来说，视觉感知可以建模为信息交互过程，而 HVS 相应地被视为一个高效的编码器或信息提取器，由此可以推导：视觉场景中包含更多信息的区域更有可能吸引视觉注意[16]，如果有关于信源的统计概率模型，那么可以使用信息论对视频内容进行量化。此外，以往的实验表明，HVS 并不能以相同的确定程度感知所有的信息内容，例如：当视频存在大量随机背景运动时，HVS 无法以感知静态图像相同的精度感知视频内容，即这种视频信号被感知时有更高的不确定度，这相当于在信息传输过程中产生了信道失真，若能将这种信道失真过程建立为与感知参量相关的模型，那么这种不确定性也能使用信息论来量化[18]。基于这个启发，依据提出的 5 个感知特征参量，提出了相应的统计概率模型，并使用信息论来量化由这些参量导致的视觉感知显著度和不确定度，并将映射到相同量纲的 5 个感知特征参量进行融合，得到时域 JND 权重模型，在空域 JND 阈值的基础上，融合时域权重得到基于空时域感知特性结合的 JND 模型。

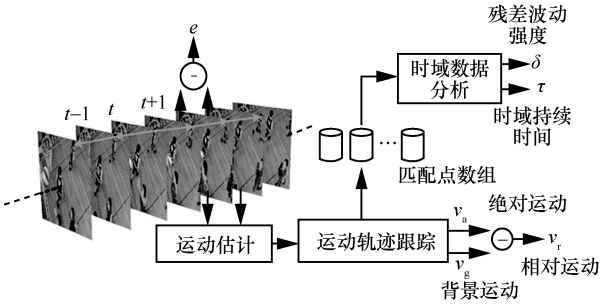


图 1　时域感知特征参量提取方法

1　感知参量同质化度量

异质感知特征参量需要寻求统一的度量方法，以有效融合其相互作用关系。从能量分配的角度探究异质感知特征参量的融合方法，基于 HVS 感知特性，提出特征参量的概率密度函数，利用感知信息论原理定量度量感知显著性和感知不确定性，实现异质感知特征参量的同质化度量。

1.1　相对运动同质化度量

一般地，运动的目标相对运动越大，运动面积越大，越能吸引人眼的注意，相应地有越大的视觉感知显著度。统计分析的结果给出了相对运动的先验概率分布，大致可以用一个幂函数来表示[17]：

$$p(v_{\mathrm{r}})=\frac{\beta_1}{v_{\mathrm{r}}^{\alpha_1}} \tag{1}$$

其中，v_r 为相对运动速度，α_1 为模型参数，β_1 为大于 0 的常量。由于相对运动的视觉感知显著度还应与运动目标的面积相关，即相对运动目标越大，越能吸引人眼关注，因此，将模型参数 α_1 建模为与运动面积相关的函数：

$$\alpha_1 = k_1 \mathrm{e}^{s(x)} \tag{2}$$

其中，$s(x)$ 为运动目标的面积，k_1 为大于零的常数，根据主观实验取值 0.2。因此，已知先验概率分布，视觉感知显著度可以使用自信息来度量：

$$I'(v_r) = -\mathrm{lb}p(v_r) = \alpha_1 \mathrm{lb}\, v_r - \mathrm{lb}\beta_1 \tag{3}$$

此外，由于人眼视觉敏感具有方向性，人眼对水平和垂直分量比较敏感，而对对角线方向不敏感，即具有倾斜效应[19]。基于此分析，考虑运动方向对视觉显著性的影响，提出基于运动方向的显著度调节因子：

$$h(\theta) = \| \varphi \times \cos(2\times\theta) \| + 1 \tag{4}$$

其中，θ 为相对运动矢量方向角，φ 为幅度调节因子，取 φ=0.3。因此，相对运动的视觉感知显著度计算式为：

$$I(v_r) = I'(v_r) \times h(\theta) \tag{5}$$

相对运动的视觉感知显著度随着运动速度和运动目标面积的增大而增大，运动矢量方向角越接近水平和垂直方向，视觉显著度越大，这符合人眼视觉特性。

1.2　背景运动同质化度量

大范围随机的背景运动会消耗人眼感知能量，降低人眼对视频细微失真的分辨能力，这种抑制效应可以认为是背景运动导致的视觉感知不确定性，等效于人眼在观察视频细节时加入了噪声。可以使用似然函数来表示这个等效噪声，参考以前的工作，这个似然函数为对数正态分布，计算式如下[17]：

$$p\left(m_1 |\ v_g\right) = \frac{1}{\sqrt{2\pi}\sigma_1 m_1} \exp\left[\frac{-\left(\mathrm{lb}m_1 - \mathrm{lb}v_g\right)^2}{2\sigma_1^2}\right] \tag{6}$$

其中，v_g 为视频背景运动，m_1 为刺激 v_g 所产生的等效噪声。高斯曲线宽度参数 $\sigma_1 = \frac{\lambda_1}{c^{\gamma_1}}$，$c$ 为对比度阈值，参数 λ_1、γ_1 是大于 0 的常数。

相邻像素运动矢量大小的不一致会对视频感知质量产生负面影响，类似于抖动现象。假设 $v_a(t,i,j)$ 为第 t 帧当前像素点的运动矢量大小，在一定区域范围 $\Omega(i,j)$ 内，定义空域运动矢量大小不规则度 Δv_s 为这一范围内所有运动矢量的标准差。因此，加入 Δv_s 调节背景运动导致的视觉感知不确定度，使用信息熵度量计算式如下：

$$U\left(v_g\right) = -\int_{-\infty}^{\infty} p\left(m_1 |\ v_g\right) \mathrm{lb}p\left(m_1 |\ v_g\right) d\mathrm{m}_1 = \mathrm{lb}v_g - \gamma_1 \mathrm{lb}c + \rho_1 \mathrm{lb}\Delta v_s + \eta_1 \tag{7}$$

其中，$\eta_1 = \mathrm{lb}\lambda_1 + \frac{1}{2} + \frac{1}{2}\mathrm{lb}(2\pi)$ 为常数，ρ_1 为调节控制参数。以上计算式计算的视觉感知不确定度与人眼的主观感知是一致的。一方面，相对于静态图像来说，运动的视频中的细微失真更难察觉，视觉感知不确定度随背景运动的增大而增大；另一方面，对比度阈值越大，人眼看得越清楚，失真更易于察觉，因此不确定度随对比度阈值的增大而减小，此外，空域运动矢量大小不规则度越大会导致更强的空域掩蔽，视觉感知不确定度越大。

1.3　时域持续时间同质化度量

人眼视觉系统具有近因效应和非对称感知的能力。视频中持续时间较长的区域部分，会在人脑中留下更强的短暂记忆效应，对于这些图像内容，人眼具有较高的感知敏感度。已有实验证明，在一定时间范围内，人眼的感知敏感度随持续时间的增加而增加，当超过这个范围，人眼感知灵敏度趋于稳定，不再受时间变化的影响[20]。时域持续时间 τ 与人眼视觉显著性 $I'(\tau)$ 的关系可以使用 sigmoid 函数来描述：

$$I'(\tau) = \frac{1}{1 + b \times \mathrm{e}^{a\times\tau}} \tag{8}$$

其中，a 和 b 均为常数调整因子，实验中取 b=1.65，a=−3.4。此外，时域持续时间相同时，观察运动的目标往往比静止的目标需要消耗人眼更多能量，因此静止的区域若产生失真波动，更能引起人眼关注。而根据主观实验发现，在大量随机运动中，失真更易于掩藏，而在规则和平稳的运动中，失真更难掩藏[21]。利用时域持续时间内目标运动方向的改变量 ω 来衡量目标运动是否规则，并结合目标运动矢量大小 v_a 调节时域持续时间的显著度：

$$I(\tau)=-\mathrm{lb}p(\tau)=\frac{1}{1+b\times \mathrm{e}^{a\times\tau}}-\alpha_2\mathrm{lb}v_a-\beta_2\mathrm{lb}\omega \quad (9)$$

其中，参数 α_2、β_2 为大于零的常数。单位时间内目标运动方向的改变量 ω 定义为 $\omega=\dfrac{\Delta\theta_t}{\tau}$，$\Delta\theta_t$ 为持续时间内运动方向角的偏转量：

$$\Delta\theta_t(t,i,j)=\sum_{t=2}^{\tau}\|\theta(t,i,j)-\theta(t-1,p,q)\|\times\kappa(t-1,p,q) \quad (10)$$

其中，$\theta(t,i,j)$ 为当前像素(i,j)的运动矢量方向角，$\theta(t-1,p,q)$ 为其在 t−1 帧中的最佳匹配位置(p,q)的运动矢量角,由于目标在运动过程中可能存在遮挡或暴露等情况，使用 $\kappa(t-1,p,q)$ 描述运动匹配点在 t−1 帧是否存在，若匹配点存在则 $\kappa(t-1,p,q)=1$，否则为 0。

式（9）计算的视觉感知显著度在一定时间阈值内随持续时间增加而增加，超过这个阈值后会趋于饱和。此外，绝对运动矢量越大，且单位时间内时域运动方向的不规则度越大，人眼视觉显著度越小，式（9）符合 HVS 特性。

1.4 残差波动强度同质化度量

运动轨迹上像素值随时间的频繁变化，表现为闪烁、蚊式噪声等，会吸引人眼关注，并使观众产生厌烦感，这相当于视频播放过程中的时域不确定度源。这种不确定性相当于在观看视频时加入了等效噪声，可以使用似然函数来表示这个等效噪声。考虑时域 HVS 特性，将概率密度函数建模为对数正态分布形式，计算式表示为：

$$p\left(m_2|\ \delta\right)=\frac{1}{\sqrt{2\pi}\sigma_2 m_2}\exp\left[\frac{-\left(\mathrm{lb}m_2-\mathrm{lb}\delta\right)^2}{2\sigma_2^2}\right] \quad (11)$$

其中，δ 为刺激源，即残差波动强度。假设当前帧$(i,\ j)$处像素点在 t−1 帧中最佳匹配点位于(p,q)，那么(i,j)前向匹配点预测误差定义为：

$$\mathrm{err}(t,i,j)=\left|\left|f(t,i,j)-f(t-1,p,q)\right|\right|\times\kappa(t-1,p,q) \quad (12)$$

其中，$f(t,i,j)$ 和 $f(t-1,p,q)$ 分别为当前帧像素点及其 t−1 帧中最佳匹配点的像素值。像素时域持续时间内，计算所有帧间匹配点的误差，得到误差矩阵，残差波动强度 δ 即为该矩阵的标准差。m_2 为加入的等效噪声，宽度参数 $\sigma_2=\lambda_2\Delta v_t^{\gamma_2}$，其中 λ_2、γ_2 是大于 0 的常数，Δv_t 为时域运动矢量的不规则度，定义为：

$$\Delta v_t(t,i,j)=\sum_{t=2}^{\tau}\|v_a(t,i,j)-v_a(t-1,p,q)\|\cdot k(t-1,p,q) \quad (13)$$

因此，加入时域运动矢量大小不规则度调节残差波动强度导致的视觉感知不确定度，使用信息熵度量计算式如下：

$$U(\delta)=-\int_{-\infty}^{\infty}p\left(m_2|\ \delta\right)\mathrm{lb}p\left(m_2|\ \delta\right)\mathrm{d}m_2=\mathrm{lb}\delta+\gamma_2\mathrm{lb}\Delta v_t+\eta_2 \quad (14)$$

其中，$\eta_2=\mathrm{lb}\lambda_2+\dfrac{1}{2}+\dfrac{1}{2}\mathrm{lb}(2\pi)$ 为一个常数。式（14）中的视觉感知不确定度受残差波动强度和时域运动矢量大小的不规则度的共同影响：随残差波动强度和时域运动矢量大小的不规则度的增大而增大。这符合人眼视觉特性。

1.5 帧间残差同质化度量

贝叶斯脑理论表明，对于当前输入的图像，人脑会自动预测后续视频帧，以实现对输入场景的感知。输入图像与人脑中的预测图像之间的误

差是不可预测部分，即时域分量的不确定性。这种不确定性相当于在观看视频时加入了等效噪声，可以使用似然函数来表示这个等效噪声，考虑时域 HVS 特性，将概率密度函数建模为对数正态分布形式，计算式表示为：

$$p\left(m_3 \mid e\right)=\frac{1}{\sqrt{2\pi}\sigma_3 m_3}\exp\left[\frac{-\left(\mathrm{lb}m_3-\mathrm{lb}e\right)^2}{2\sigma_3^2}\right] \tag{15}$$

其中，e 为刺激源，即相邻帧间预测残差，计算式为[22]：

$$e(t,i,j)=\frac{f(t,i,j)-f(t-1,i,j)+g(t,i,j)-g(t-1,i,j)}{2} \tag{16}$$

其中，$f(t,i,j)$ 和 $f(t-1,i,j)$ 分别为 t 帧和 $t-1$ 帧中对应亮度值之差，$g(t,i,j)$ 和 $g(t-1,i,j)$ 为相应平均背景亮度值之差。m_3 为加入的等效噪声。宽度参数 σ_3 对 e 来说，可以认为是一个常数，但受亮度自适应（LA）阈值的影响。LA 表示人眼对不同背景亮度的敏感度，主观实验表明，HVS 对过高或过低的背景亮度区域不敏感，而对中等亮度的区域高度敏感。LA 越大，人眼对亮度变化的感知阈值越大，因此，曲线宽度，$\sigma_3=\lambda_3\mathrm{LA}^{\gamma_3}$。其中，$\lambda_3$、$\gamma_3$ 是大于 0 的常数。相邻帧间预测残差导致的视觉感知不确定度可以用信息熵来度量：

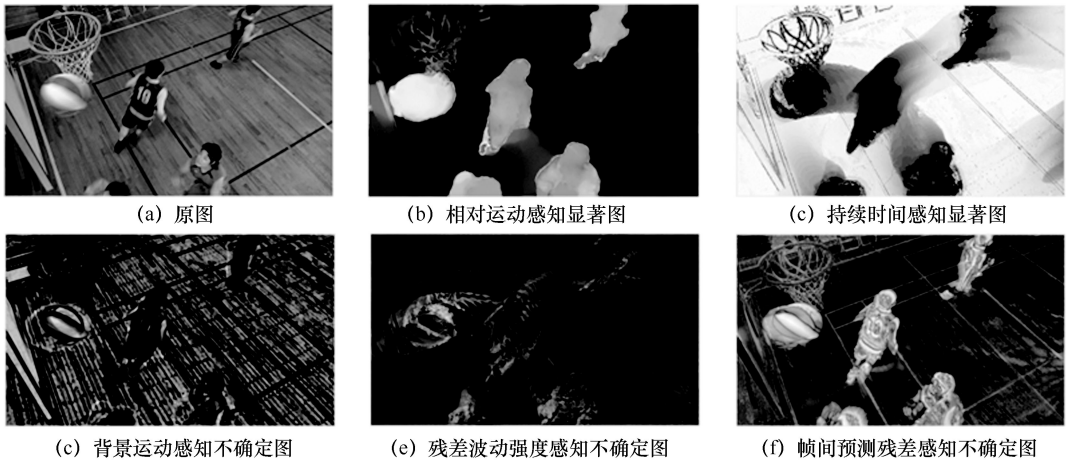
$$U(e)=-\int_{-\infty}^{\infty}p\left(m_3\mid e\right)\mathrm{lb}p\left(m_3\mid e\right)\mathrm{d}m_3=\mathrm{lb}e+\gamma_3\mathrm{lbLA}+\eta_3 \tag{17}$$

其中，$\eta_3=\mathrm{lb}\lambda_3+\frac{1}{2}+\frac{1}{2}\mathrm{lb}(2\pi)$ 为常数。式（17）中的视觉感知不确定度受相邻帧间预测残差和亮度自适应阈值的共同影响：随残差波动强度和亮度自适应阈值的增大而增大，这符合人眼视觉特性。

以上计算式中常量的取值参考文献[16]中的方法，均为根据主观实验手动选取，常量的小范围波动对最终结果影响不大。图 2（a）为“BasketBallDrill”序列的第 5 帧的原始图像，图 2（b）和图 2（c）分别为相对运动和时域持续时间产生的感知显著度图，越亮的区域表示有更大的视觉感知显著度；图 2（d）、图 2（e）和图 2（f）分别为背景运动、运动轨迹上残差波动强度和相邻帧间预测误差产生的感知不确定度图，越亮的区域对应更大的视觉感知不确定度。

2　时域参量融合

由于已经计算了相对运动、时域持续时间等特征参量引起的视觉感知显著度，背景运动、时域轨迹上的残差波动强度、相邻帧间预测残差引起的感知不确定度。基于自信息和信息熵的理论，将 5 个异质感知特征参量同质化到统一尺度，并



(a) 原图　(b) 相对运动感知显著图　(c) 持续时间感知显著图

(c) 背景运动感知不确定图　(e) 残差波动强度感知不确定图　(f) 帧间预测残差感知不确定图

图 2　时域特征参量的感知显著图和感知不确定图

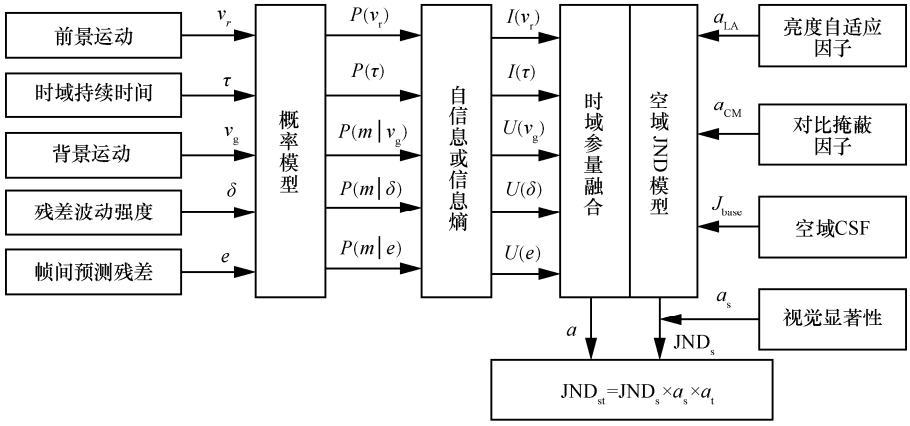


图 3　空时域 JND 模型框架

考虑视觉显著性的影响，加入显著性调节因子，对 JND 进行时域和空域调节，空时域 JND 模型框图如图 3 所示。

空时域 JND 阈值 $\mathrm{JND}_{\mathrm{ST}}(t,n,i,j)$ 计算如下：

$$\mathrm{JND}_{\mathrm{ST}}(t,n,i,j)=\mathrm{JND}_{\mathrm{s}}(t,n,i,j)\cdot\alpha_t(t,n)\cdot\alpha_{\mathrm{S}}(t,n) \quad (18)$$

其中，$\mathrm{JND}_{\mathrm{s}}(t,n,i,j)$ 表示第 t 帧的第 n 个特定大小图像块中坐标为 (i,j) 像素的空域 JND 值，考虑了亮度自适应、对比掩蔽、空域对比敏感函数等[9]：

$$\begin{aligned}&\mathrm{JND}_{\mathrm{s}}(t,n,i,j)=\\&J_{\mathrm{base}}(t,n,i,j)\times a_{\mathrm{LA}}(t,n,i,j)\times a_{\mathrm{CM}}(t,n,i,j)\end{aligned} \quad (19)$$

其中，$J_{\mathrm{base}}(t,n,i,j)$ 为只考虑空域 CSF 的基本 JND 阈值，$a_{\mathrm{LA}}(t,n,i,j)$ 和 $a_{\mathrm{CM}}(t,n,i,j)$ 为基于背景亮度和对比度建模的亮度自适应因子和对比掩蔽因子。

在空域 JND 阈值的基础上，加入时域调节因子 $a_{\mathrm{t}}(t,n)$ 和显著性调节因子 $a_{\mathrm{s}}(t,n)$，以获得更加准确的感知失真阈值，表示为第 t 帧的第 n 个特定大小图像块内时域调整因子 a_{t} 和空域显著度调整因子 a_{s} 的均值。为融合 5 个特征参量的视觉感知显著度和不确定度，使用 NAMM 模型分别得到显著度 I_{t} 和不确定度 U_{t}：

$$\begin{aligned}&I_{\mathrm{t}}=\mathrm{I}(v_{\mathrm{r}})+I(\tau)-0.3\times\min(\mathrm{I}(v_{\mathrm{r}}),I(\tau)),\\&U_{\mathrm{t}}=U(v_{\mathrm{g}})+U\left(e\right)+U(\delta)-\\&0.3\times\min(U(v_{\mathrm{g}}),U(e),U(\delta))\end{aligned} \quad (20)$$

由于需要寻求 U_{t}、I_{t} 和 $a_{\mathrm{t}}(t,n)$ 的关系，根据 HVS 特性，I_{t} 越大，感知显著度越大；U_{t} 越小，感知不确定度越小，那么 $U_{\mathrm{t}}-I_{\mathrm{t}}$ 大于零时对应的感知不确定度越大，JND 阈值越大；$U_{\mathrm{t}}-I_{\mathrm{t}}$ 小于零时对应的感知显著度越大，JND 阈值越小。因此，sigmoid 函数比较适合用来模拟这种关系，使用如下计算式来计算时域权重：

$$a_{\mathrm{t}}=\frac{1}{1+\mathrm{e}^{-\mu\times(U_{\mathrm{t}}-I_{\mathrm{t}})}}+\frac{1}{2} \quad (21)$$

其中，μ 控制曲线斜率，根据主观实验 $\mu=6$。显著性调节因子 a_{s} 为利用 Wu 等人[23]提出的显著性检测方法计算得到的像素显著度值 s_{a} 后，利用计算式：

$$a_{\mathrm{s}}=1-\left(s_{\mathrm{a}}-\phi_1\right)\times\phi_2 \quad (22)$$

其中，ϕ_1、ϕ_2 为大于零的常量，实验中取 $\phi_1=0.25$，$\phi_2=0.55$。其中，s_{a} 被归一化到 0 至 1 之间。当 s_a 越接近 1，表明显著度越高，因此对应的 JND 阈值越小。

为了获得直观的效果，将 a_{t} 映射到[0,1]，图 4（a）和图 4（b）分别表示得到的空域显著图和时域权重图，图 4（c）为最终的空时域 JND 图。图 4（a）中，越亮的区域表示视觉显著度越大，图 4（b）中，越亮的区域表示视觉感知不确定度越大，图 4（c）中越亮的区域表示能够容纳的噪声越多，JND 阈值越大。

3　实验结果与分析

为了验证 JND 模型的有效性，分别进行了客

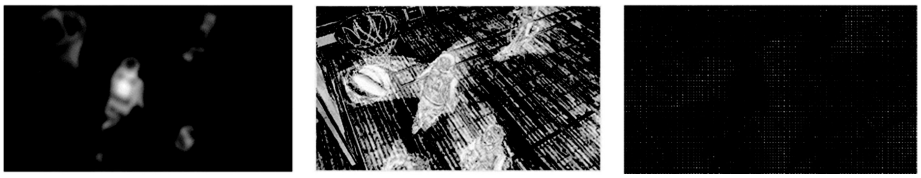
观和主观视频质量评价实验。选取 HEVC 测试序列中的 9 个视频进行测试。其中有 5 个视频分辨率为 1 920 dpi×1 080 dpi 的全高清分辨率的视频，分别为 BasketballDrive(Ba)、BQTerrance(BT)、Cactus(CT)、Kimonol(Ki)和 ParkScence(Pa)，另外 4 个视频 BasketballDrill(BD)、BQMall(BM)、PartyScence(PS)和 RaceHorse(RH)为分辨率 832 dpi×480 dpi 的视频。根据计算式：

$$c^{\wedge}(i,j)=c(i,j)+\varsigma\times\mathrm{rand}(i,j)\times\mathrm{JND}_{\mathrm{ST}}(i,j) \quad (23)$$

向视频每帧的 DCT 系数加入噪声。其中，$c(i,j)$ 和 $c^{\wedge}(i,j)$ 分别为 (i,j) 位置的 DCT 系数及其失真系数的值，$\mathrm{rand}(i,j)$ 随机取值+1 或−1，ς 控制加入噪声的能量大小。分别计算失真视频的峰值信噪比[24]（peak signal-to-noise ratio，PSNR），PSNR 越大，表示加入噪声越多。

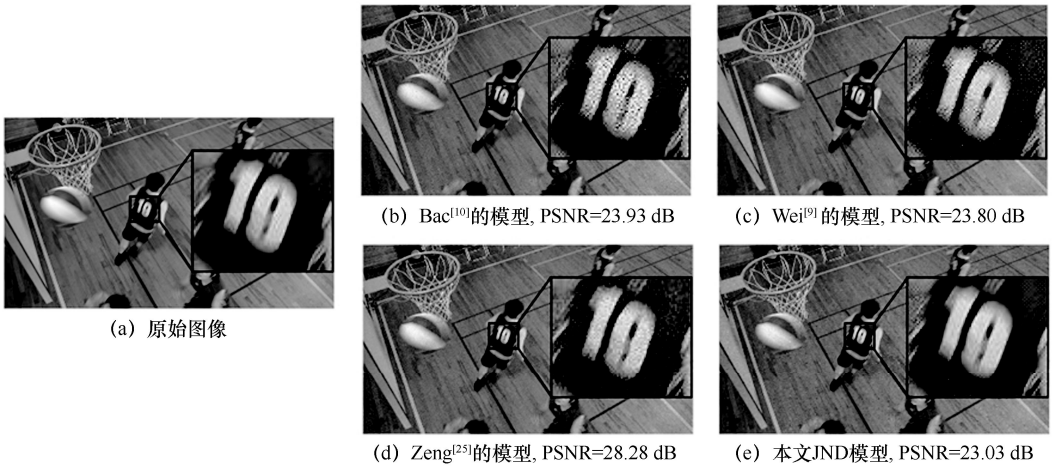
图 5 为不同 JND 模型处理结果，表示“BasketballDrill”序列第 5 帧根据不同 JND 模型加入噪声后的图像，图 5 中被放大的细节部分为人眼更易注意的区域。4 个 JND 模型分别为 Bae[10]的模型、Wei[9]的模型、Zeng[25]的模型，以及所提出的 JND 模型，相应的 PSNR 分别为 23.93 dB、23.80 dB、28.28 dB 以及 23.03 dB。由于放大的区域对于整幅图像来说是显著的，因此，这个区域 JND 阈值较低，能加入的噪声较少。而从图 5 可以看出，图 5（c）和图 5（d）中运动员身上都有明显可见的噪声，且图 5（b）球衣上的数字模糊了，而提出的模型考虑了这一特性，计算所得这部分区域的 JND 阈值更小，从图 5 中可以看出图 5（e）与其余 3 张图像相比有更好的感知质量，且 PSNR 值最低，为 23.03 dB。因此，以上客观和主观评估都证明了提出的调节因子是有效的，能够防止 JND 阈值被过高估计。

图 6 为分别利用 Bae[10]、Wei[9]、Zeng[25]和提出的 JND 模型向 Basketball 序列加入噪声后每一帧的 PSNR 曲线。由图 6（b）可以看出 Wei 提出的模型的相邻帧 PSNR 变化波动较小，JND 阈值受时域影响不大。此外，虽然 Wei 提出的模型的时域调节因子考虑了运动的方向性，但其假设所



（a）空域显著图　（b）时域JND权重系数图　（c）空时域JND图

图 4　空时域 JND 结果



（a）原始图像　（b）Bac[10]的模型, PSNR=23.93 dB　（c）Wei[9]的模型, PSNR=23.80 dB

（d）Zeng[25]的模型, PSNR=28.28 dB　（e）本文JND模型, PSNR=23.03 dB

图 5　不同 JND 模型处理结果

有的运动目标都被人眼跟踪，导致不被跟踪的目标 JND 阈值被高估。Bae 充分考虑了人眼的运动的特性，并结合了中心凹掩蔽效应，有较好的效果，但忽略了运动的视频连续帧之间残差导致的掩蔽效应，这会影响时域 JND 阈值的估计。图 6（c）中 Zeng 主要基于区域方向对纹理掩蔽效果进行了改善，而没有考虑时域掩蔽效应，相邻帧间JND阈值不受时域变化的影响，因此PSNR波动较小。图 6（d）充分考虑了视频时域特征参量对 HVS 特性的影响，并在信息论的基础上将异质感知特征参量进行融合，保证了更好的融合效果。由于时域掩蔽的加入，会影响相邻帧之间加入噪声的总量，图 6 中的 PSNR 曲线间接表明，JND 阈值有了显著调整。

由于视频的最终接收者为人眼，仅仅依赖 PSNR 值无法衡量视频质量的好坏，因此采用视频多方法评估融合（video multimethod assessment fusion，VMAF）得分和平均主观得分差（differential mean opinion score，DMOS）来评价视频的主观质量。利用双刺激连续质量标度方法（double stimulus continuous quality scale，DSCQS）进行主观质量评估实验，实验过程中邀请了 14 位视力正常或佩戴视力矫正器的评估者，显示器的分辨率为 1 920 dpi×1 080 dpi，并将观看距离设置为显示屏高度的 3 倍。图 7 展示了 DSCQS 的使用方法，失真视频 A 和原始视频 B 随机呈现在观察者眼前，在交替显示两次或多次后，观看者对两个视频分别进行评分，用 DMOS 值来衡量失真视频与原始视频的主观质量接近程度，计算式如下：

$$\mathrm{DMOS} = \mathrm{MOS}_{\mathrm{ori}} - \mathrm{MOS}_{\mathrm{jnd}} \tag{24}$$

$\mathrm{MOS}_{\mathrm{ori}}$ 表示原始视频的得分，$\mathrm{MOS}_{\mathrm{jnd}}$ 表示根据 JND 模型加噪声后失真视频的得分，评分标准为非常好（80～100）、好（60～80）、还可以（40～60）、差（20～40）、非常差（0～20），两者差值越小，表示失真视频和原始视频的主观质量越接近。因此，性能更好的JND模型应该表现为失真视频有更低的 PSNR 同时有更高的 VMAF 得分且 DMOS 值相近或更小。

表 1 对比了由 4 个 JND 模型向不同分辨率大小的视频加入噪声后的 PSNR、VMAF 得分和相应的 DMOS 值。基于 Wei、Bae、Zeng 以及所提出的 JND 模型加入噪声后的平均 PSNR 分别为 27.68 dB、28.16 dB、32.27 dB 以及 26.93 dB，相应的平均 VMAF 值为 98.71、97.88、97.16、99.23，平均 DMOS 值分别为 14.34、16.04、19.62、11.66。PSNR 越小表示加入噪声越多，DOMS 值越小表

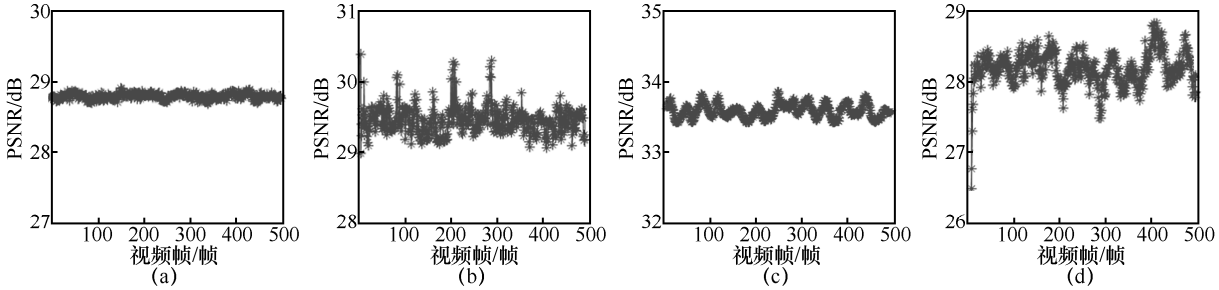


图 6　根据不同 JND 模型向 Basketball 序列加入噪声后 PSNR 曲线

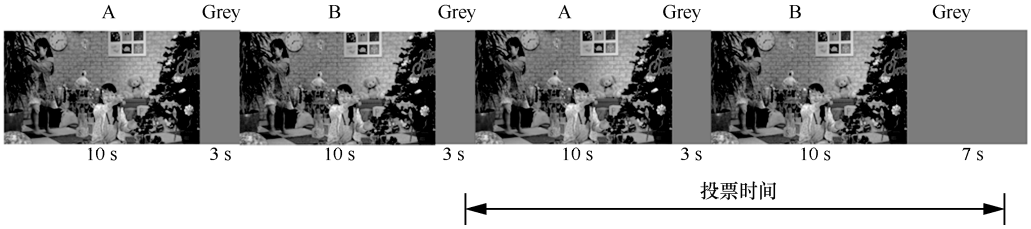


图 7　DSCQS 主观评价方法

表 1　不同 JND 模型性能对比

测试序列	Wei[9]提出的模型			Bae[10]提出的模型			Zeng[25]提出的模型			本文所提模型		
	PSNR	VMAF	DMOS	PSNR	VMAF	DMOS	PSNR	VMAF	DMOS	PSNR	VMAF	DMOS
Ba	26.66	99.90	12.67	27．09	99.76	15.91	33.53	99.80	19.52	25.88	99.98	9.33
BT	26.48	99.17	16.50	25.76	98.26	19.65	34.34	99.61	18.39	25.40	99.31	12.89
CT	28.20	97.36	11.95	28.73	96.88	11.67	31.33	94.35	22.58	27.89	98.78	7.38
Ki	27.50	98.20	15.58	27.55	95.49	22.20	29.92	94.27	22.67	27.11	98.90	13.32
Pa	28.21	97.56	16.89	28.20	95.37	18．11	30.41	93.52	20.12	27.69	98.26	15.11
BD	28.80	99.76	14.33	29.53	99.33	12.83	33.61	98.59	19.12	28.17	99.95	11.50
BM	28.16	99.76	12.67	30.00	99.84	10.78	32.09	99.15	17.33	26.43	99.97	9.33
PS	28.26	96.72	13.33	28.72	96.08	15.00	32.62	95.14	19.65	27.91	97.94	11.20
RH	26.88	99.98	15.17	27.88	99.90	18.24	32.56	99.96	17.17	25.89	99.98	14.89
平均值	27.68	98.71	14.34	28.16	97.88	16.04	32.27	97.16	19.62	26.93	99.23	11.66

示加入噪声的视频主观感知质量越接近于原始视频，有相对更小的感知失真。由此可见，在加入噪声更多的情况下，提出的 JND 模型与其他模型相比有更好的感知质量，另外 VMAF 指标分数可以看出，由提出的 JND 模型加入噪声后的视频质量更好的情况下，PSNR 至少降低了 0.75 dB。因此可以得出结论：与其他模型相比，所提出的 JND 模型能够隐藏更多的噪声，同时使视频拥有更好的感知质量。此外，实验过程中 4 个模型的 S 值分别为 0.9、0.6、0.8 和 1，进一步证明提出的模型估计 JND 阈值更加准确。

4　结束语

本文通过探索时域 HVS 特性，分析视频目标时域的运动特征，提出了一个新的时域 JND 权重模型。将相对运动、运动轨迹上的持续时间等注意力激励源，背景运动、运动轨迹上残差波动强度、相邻帧间预测残差等注意力不确定源，利用统计信息论原理映射到信息量度量的特征空间，采用自信息度量视觉感知显著度，信息熵度量感知不确定度。最后，将映射到统一尺度上的 5 个感知特征参量进行融合，得到时域权重因子，并考虑空域视觉显著性的影响，在空域 JND 阈值的基础上加入显著性调节因子和时域权重因子，以此构建了适用于视频的空时域 JND 模型。实验结果证明，提出的模型性能优于现有 JND 模型。

参考文献：

[1] YUAN D, ZHAO T S, XU Y W, et al. Visual JND: a perceptual measurement in video coding[J]. IEEE Access, 2019(7): 29014-29022.

[2] KORHONEN J. Two-level approach for no-reference consumer video quality assessment[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2019, 28(12): 5923-5938.

[3] CHOU C H, LI Y C. A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile[J]. IEEE Transactions on Circuits and Systems for Video Technology, 1995, 5(6): 467-476.

[4] YANG X K, LING W S, LU Z K, et al. Just noticeable distortion model and its applications in video coding[J]. Signal Processing: Image Communication, 2005, 20(7): 662-680.

[5] CHIN Y J, BERGER T. A software-only videocodec using pixelwise conditional differential replenishment and perceptual enhancements[J]. IEEE Transactions on Circuits and Systems for Video Technology, 1999, 9(3): 438-450.

[6] KELLY D H. Motion and vision. II. Stabilized spatio-temporal threshold surface[J]. Journal of the Optical Society of America, 1979, 69(10): 1340-1349.

[7] DALY S. Engineering observations from spatiovelocity and spatiotemporal visual models [M]. Vision Models and Applica-

tions to Image and Video Processing. Heidelberg: Springer, 2001: 179-200.

[8] JIA Y, LIN W, KASSIM A A. Estimating just-noticeable distortion for video[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2006, 16(7): 820-829.

[9] WEI Z Y, NGAN K N. Spatio-temporal just noticeable distortion profile for grey scale image/video in DCT domain[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2009, 19(3): 337-346.

[10] BAE S H, KIM M. A DCT-based total JND profile for spatiotemporal and foveated masking effects[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(6): 1196-1207.

[11] WANG H Q, GAN W H, HU S D, et al. MCL-JCV: a JND-based H.264/AVC video quality assessment dataset[C]//Proceedings of 2016 IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2016:1509-1513.

[12] WANG H Q, KATSAVOUNIDIS I, ZHOU J T, et al. VideoSet: a large-scale compressed video quality dataset based on JND measurement[J]. Journal of Visual Communication and Image Representation, 2017, 46: 292-302.

[13] KI S, BAE S H, KIM M, et al. Learning-based just-noticeable-quantization- distortion modeling for perceptual video coding[J]. IEEE Transactions on Image Processing, 2018, 27(7): 3178-3193.

[14] LIU H H, ZHANG Y, ZHANG H, et al. Deep learning-based picture-wise just noticeable distortion prediction model for image compression[J]. IEEE Transactions on Image Processing, 2020, 29: 641-656.

[15] ZHANG Y, LIU H H, YANG Y, et al. Deep learning based just noticeable difference and perceptual quality prediction models for compressed video[J]. IEEE Transactions on Circuits and Systems for Video Technology, 6224(99): 1.

[16] WANG Z, LI Q. Video quality assessment using a statistical model of human visual speed perception[J]. Journal of the Optical Society of America A, Optics, Image Science, and Vision, 2007, 24(12): B61-B69.

[17] MACKNIK S L, LIVINGSTONE M S. Neuronal correlates of visibility and invisibility in the primate visual system[J]. Nature Neuroscience, 1998, 1(2): 144-149.

[18] STOCKER A A, SIMONCELLI E P. Noise characteristics and prior expectations in human visual speed perception[J]. Nature Neuroscience, 2006, 9(4): 578-585.

[19] PETERSON H A, AHUMADA A J J, WATSON A B. Improved detection model for DCT coefficient quantization[C]// Proceedings of the Human Vision, Visual Processing, and Digital Display IV, F, 1993. [S.l.:s.n.], 1993: 191-201.

[20] BARKOWSKY M, BIALKOWSKI J, ESKOFIER B, et al. Temporal trajectory aware video quality measure[J]. IEEE Journal of Selected Topics in Signal Processing, 2009, 3(2): 266-279.

[21] HU S D, JIN L N, WANG H L, et al. Objective video quality assessment based on perceptually weighted mean squared error[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(9): 1844-1855.

[22] CHEN Z, GUILLEMOT C. Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable-distortion model[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2010, 20(6): 806-819.

[23] WU J, QI F, SHI G, et al. Non-local spatial redundancy reduction for bottom-up saliency estimation [J]. Journal of Visual Communication and Image Representation, 2012, 23(7): 1158-1166.

[24] 崔帅南, 彭宗举, 邹文辉, 等. 多特征融合的合成视点立体图像质量评价[J]. 电信科学, 2019, 35(5): 104-112.
CUI S N, PENG Z J, ZOU W H, et al. Quality assessment of synthetic viewpoint stereo image with multi-feature fusion[J]. Telecommunications Science, 2019, 35(5): 104-112.

[25] ZENG Z P, ZENG H Q, CHEN J, et al. Visual attention guide dpixel-wise just noticeable difference model[J]. IEEE Access, 2019, 7: 132111-132119.

[作者简介]

邢亚芬（1997− ），女，杭州电子科技大学硕士生，主要研究方向为感知视频编码。

殷海兵（1974− ），男，博士，杭州电子科技大学教授，主要研究方向为数字视频编解码。

王鸿奎（1990− ），男，华中科技大学博士生，主要研究方向为感知视频编码。

骆琼华（1998− ），女，杭州电子科技大学硕士生，主要研究方向为感知视频编码。